



**caBIG**

*cancer Biomedical  
Informatics Grid*



# **A Practical Guide to CDEs and caDSR**

August 24, 2004

Tommie Curtis

SAIC, Contractor for NCICB

# Agenda

2

- ▶ Introduction – Common Data Elements
- ▶ Governance of CDEs
- ▶ Developing CDEs
- ▶ Registering UMLs in the caDSR
- ▶ Linking to EVS
- ▶ Obtaining CDEs
- ▶ Questions and Answers

# What is a Common Data Element (CDE)?

3

A Data Element is

- a unit of data for which definition, identification, representation, and permissible values are specified by means of a set of attributes; the smallest unit of data.

A Common Data Element is

- a unit of data that has been identified for general usage; may be a data standard.

## Guiding Principles for Common Data Elements:

- Developed in partnership.
- Consensus-based.
- Leveraging national and international standards.

# What is the Purpose of a CDE?

4

*The purpose of a data element definition is to define a data element with words or phrases that describe, explain, or make definite and clear its meaning.*

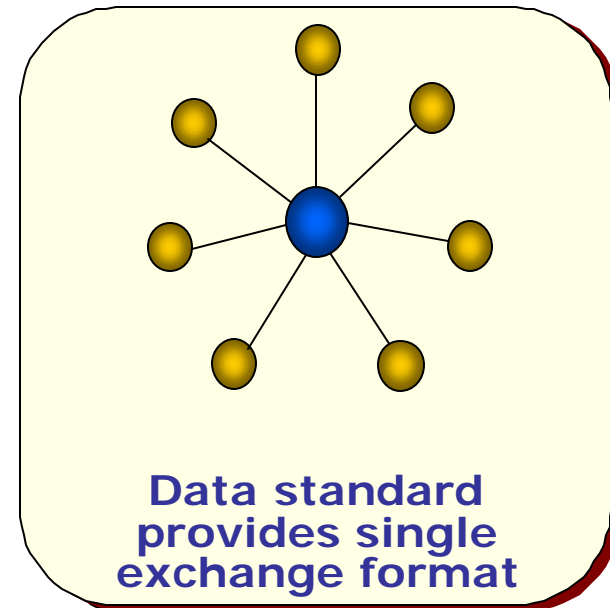
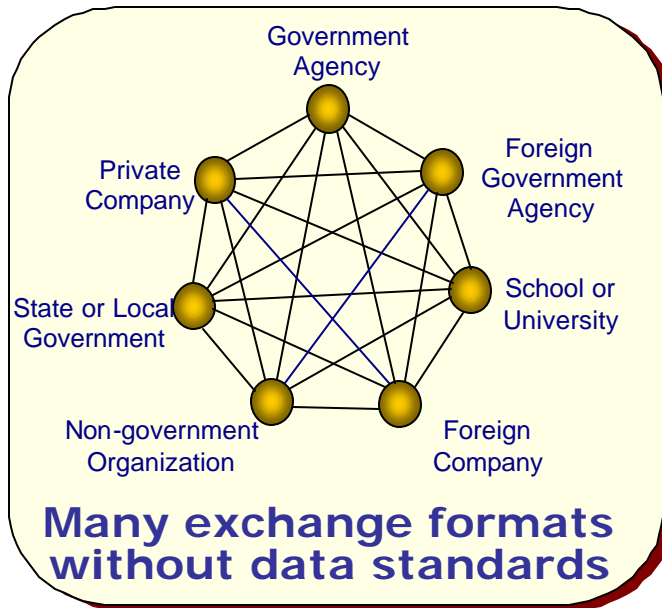
# What are Benefits of Using Documented CDEs?

5

- ▶ Facilitates common data collection by defining content and scope.
- ▶ Supports semantic data relationships.
- ▶ Defines valid values for enumerated data.
- ▶ Improves understanding of data.
- ▶ Simplifies and documents data analysis.
- ▶ Provides historical context for data collections.
- ▶ Encourages reuse of existing data structures.

# Why are CDEs Important?

6



- ▶ CDEs help to:
  - Coordinate data development, use, sharing (exchange), and dissemination.
  - Provide accurate and precise (meaningful) data.
  - Facilitate information collection.

# What Standards Do and Don't Do?

7

## ▶ Standards Do:

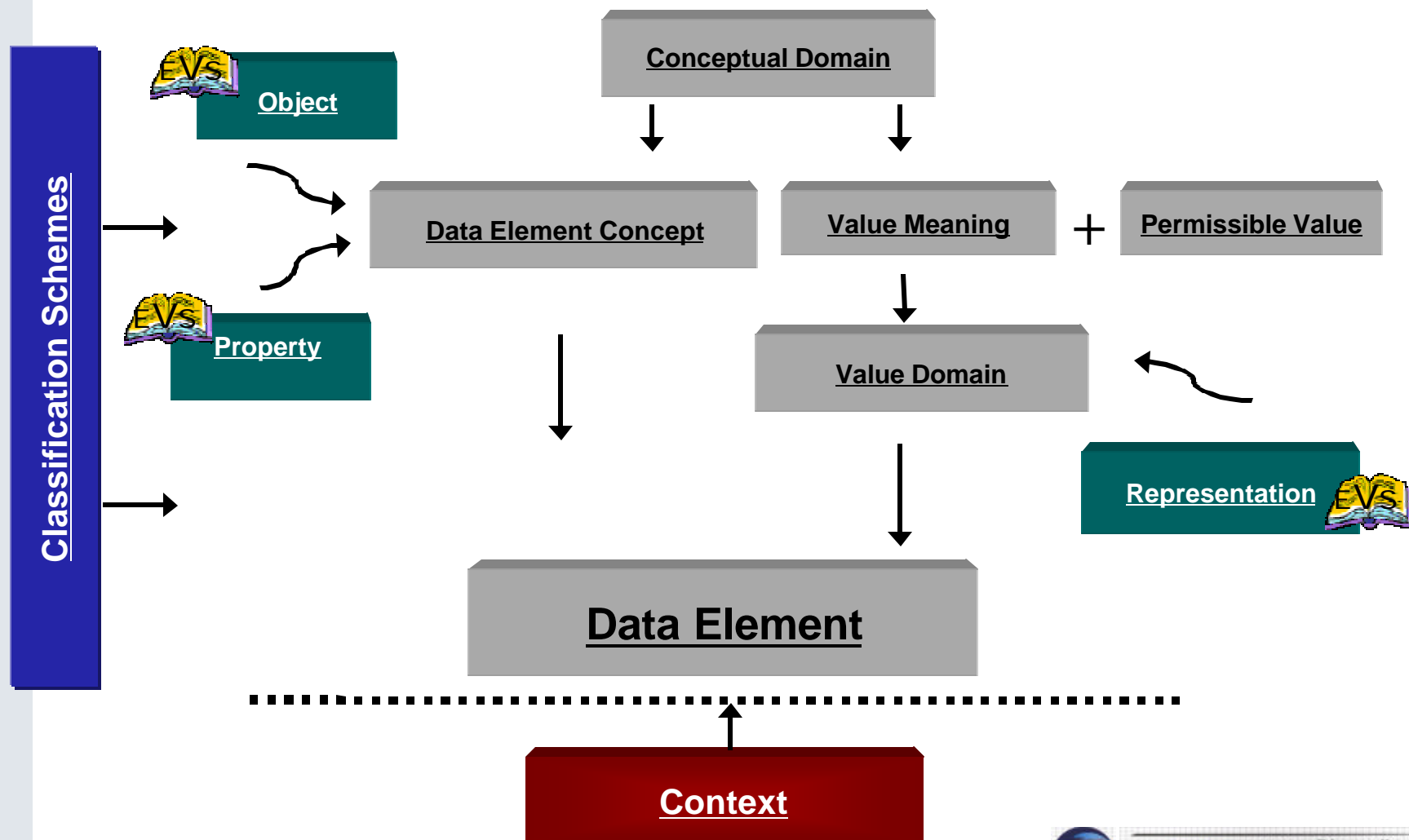
- Share data of common interest.
- Establish standard definitions and acceptable values.
- Provide a standardized set of information about the data.

## ▶ Standards Do Not:

- Tell people what data they can and cannot collect.
- Tell people how to define their unique data needs.
- Solve data quality problems.
- Establish reporting requirements.

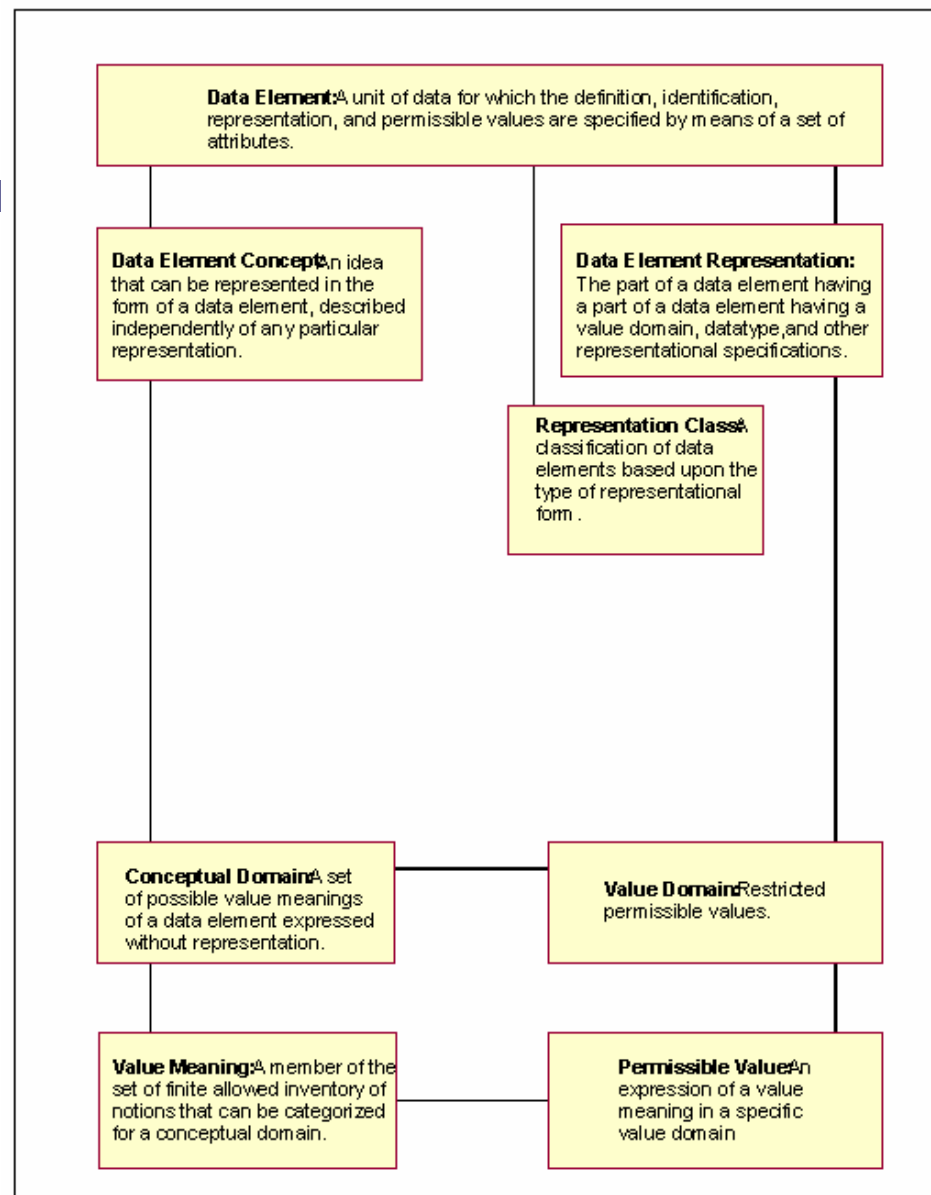
# ISO/IEC 11179 Information technology – Metadata registries, Part 3: Registry metamodel and basic attributes

8





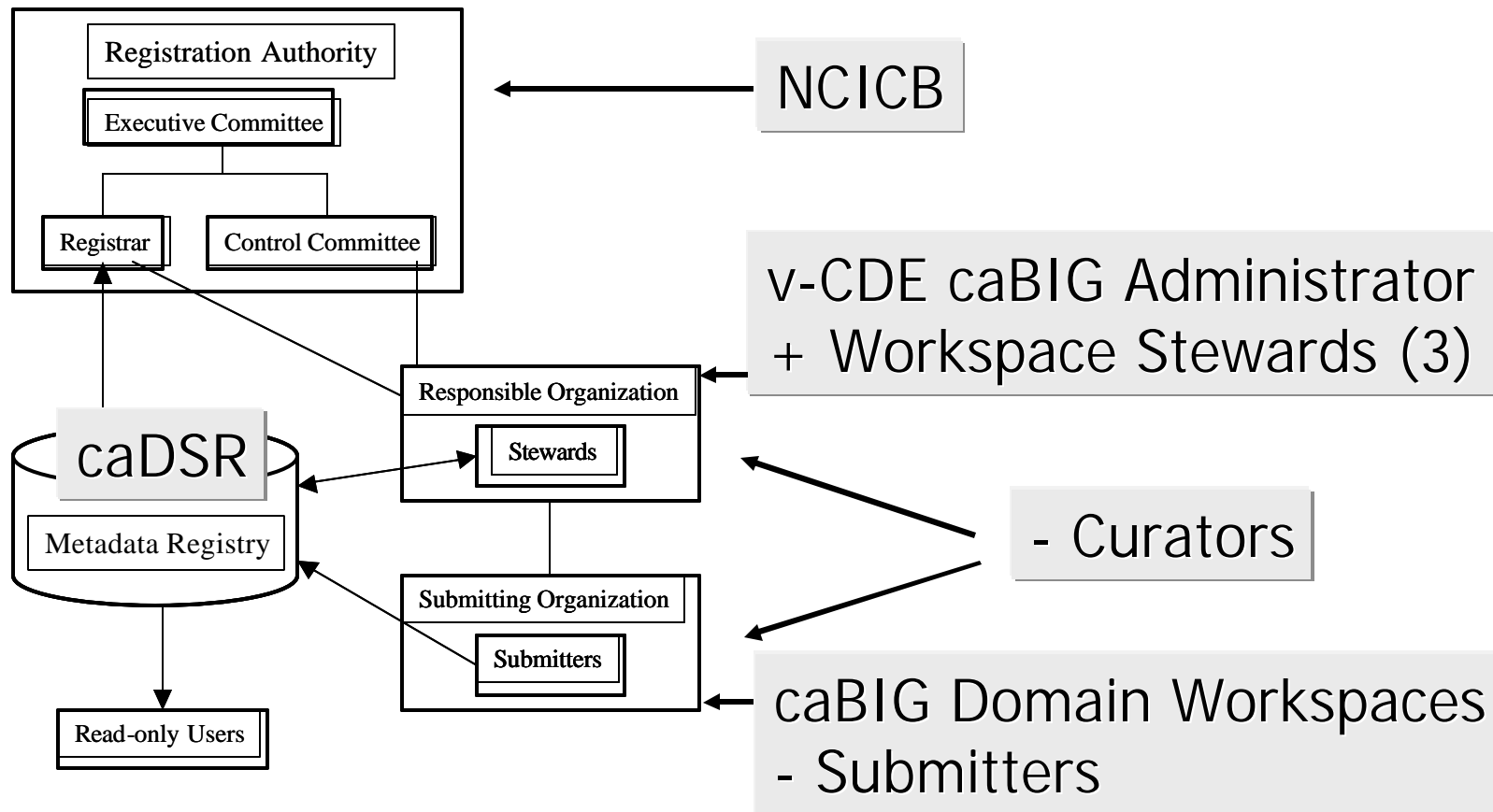
# Terms and Definitions for ISO/IEC 11179-3



# Guidance for caBIG/NCI Organizational Roles

10

## ISO 11179 Part 6: Registration of Data Elements



Organizational Roles to the Metadata registry and their relationships.

# Submitter\* Role

11

- ▶ Coordinate identification and documentation of Project CDEs
  - Propose new CDEs to support Project.
  - Ensure the quality of metadata attributes, reusing standardized data where applicable.
  - Propose workflow status.
  - Ensure that Workspace procedures and policies are followed.

\* ISO 11179 Part 6, Responsibilities of a Steward

# Curator\* Role

12

- ▶ Coordinate identification and documentation of Workspace Project CDEs
  - Ensure that appropriate CDEs are properly created.
  - Coordinate with other Workspaces Projects to attempt to prevent or resolve duplication in defining CDEs.
  - Review and attempt to resolve conflicts within their Workspace.
  - Ensure the quality of metadata attributes, reusing standardized data where applicable.
  - Propose workflow status.
  - Ensure that Workspace procedures and policies are followed.
  - Recommend CDE projects within their Workspace.

\* ISO 11179 Part 6, Responsibilities of a Steward

# Workspace Steward\* Role

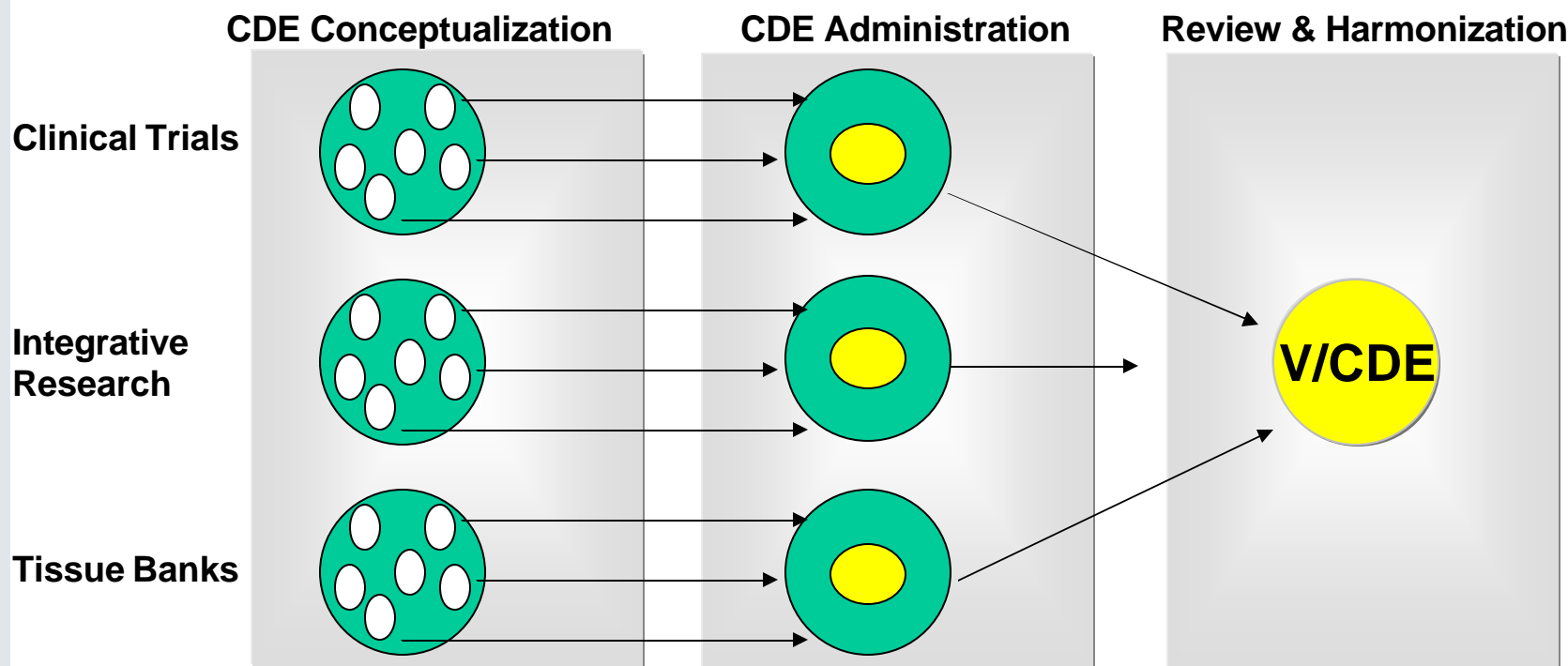
13



- ▶ Integrity, accuracy and currency of owned data elements
  - Advise workspace curators on the semantics, name, and permissible values.
  - Make decisions based on domain expertise when needed
    - Review and Approve CDEs.
    - Help curators chose between similar or like concepts.
    - Clarify the meaning of required data elements.
    - Identify pertinent workspace specific standards.
    - Review existing CDE naming conventions and adopt or modify as necessary.

\* ISO 11179 Part 6, Responsibilities of a Responsible Organization (RO)

# Scenario 4: Recommended CDE Development Model

14



-  Domain Workspace
-  Vocabularies and Common Data Elements Workspace

**Scenario 4.** Cancer centers and scientists are responsible for proposing new CDE concepts in the CDE development process. CDE administration is then carried out by a designated group from the respective Domain Workspaces, in conjunction with the V/CDE Workspace. The final review, harmonization and acceptance is carried out by the V/CDE Workspace, along with the NCI.

# What Does All This Really Mean?

15

- ▶ Workspaces will propose CDEs that meet the needs of the project teams. This development should be based on reuse of existing CDEs when possible. Communication within a workspace is important to prevent duplication of efforts.
- ▶ Workspace Steward will monitor the activities of the V-CDE workspace to be aware of CDE development efforts in other workspaces and communicate with other workspace stewards if possible duplication of efforts are occurring.
- ▶ V-CDE will review all activities and initiate harmonization activities where needed to ensure that all members of the caBIG team can share information effectively and efficiently.

# Types of Data Elements and Formats

16

- ▶ Data Elements
  - Logical – a list of the information of interest.
  - Physical – data elements as used in an application.
- ▶ Formats
  - Storage.
  - Display.
  - Exchange.



# First Steps for Development of CDEs

17

- ▶ Develop a priority list of information areas for consideration, such as lab test method identification, gene identification, data quality parameters, reagent identification, XML tags, etc.
- ▶ Identify systems that are currently being used and what information being collected for the priority information areas.
- ▶ Compile a list of applicable information standards.
- ▶ Develop use cases for data collection and analysis.
- ▶ Determine common information and permitted values.

# Example - How to Identify a Person?

18

- ▶ Person Full Name (might include Name Prefix, Person First Name, Person Middle Name or Person Initial Text, Person Last Name or Surname, Name Suffix).
- ▶ Search the caDSR to see if there are reusable CDEs (reuse is called designation)
  - Person Name (length: 35, datatype: Alphanumeric, workflow: Released).
  - Person First Name, Person Middle Name, Person Family Name, and Person Name Prefix Name or Code are also available.
- ▶ If the current CDEs do not meet your needs, contact the owning context to see if they will modify the CDE to meet your needs before creating a new CDE. A new CDE could be created by concatenation of the building blocks that make up a name.

# Naming Guidance

19

- ▶ A CDE Name should:
  - include the Object Class, Property, and Representation.
  - be definitive for value domains.
- ▶ Develop semantic rules for the source and content of words used in a name.
- ▶ Formulate syntax rules for required word order.
- ▶ Develop lexical rules covering controlled word lists, name length, character set, and language.
- ▶ Set guidelines on uniqueness of names in context.

# Lexical Principles

20

- ▶ Establish:
  - Preferred and non-preferred terms.
  - Synonyms.
  - Abbreviations.
  - Component length.
  - Spelling.
  - Permissible character set.
  - Case sensitivity.

# Using a Thesaurus

21

- ▶ Source of name components.
- ▶ Provides semantic linking of preferred terms.
- ▶ Gives guidance in using homographs.
- ▶ Shows equivalence, hierarchy, and association.
- ▶ Allows a controlled vocabulary.

# Representation Term

22

- ▶ Describes the form of the set of valid values for a data element.
- ▶ Describes the form of the representation of a data element.
- ▶ If the representation is redundant with the property term, one term or part of a term may be removed.

# Representation Terms

23

- ▶ Amount.
- ▶ Code.
- ▶ Count.
- ▶ Date.
- ▶ Group.
- ▶ Measure.
- ▶ Name.
- ▶ Number.
- ▶ Quantity.
- ▶ Rate.
- ▶ Text.
- ▶ Time.

# Qualifier Term

24

- ▶ A word or words which help define and differentiate a name within the database.
- ▶ May be attached to any basic data element name and representation class term.
- ▶ May be derived from structured sets of terms.
- ▶ Should not be redundant.



# Creating CDE Definition

25

- ▶ A data definition shall be:
  - Unique.
  - Singular.
  - A statement of concept, not its negative.
  - A descriptive phrase or sentence.
  - Commonly understood abbreviations.
  - Without embedded definitions.

*Good definitions promote the standardization and reuse of data elements, leading to data sharing and integration of information systems.*

# Data Definition Guidelines

26

- ▶ State the essential meaning of the concept.
- ▶ Be precise and unambiguous.
- ▶ Be concise.
- ▶ Be able to stand alone.
- ▶ Be expressed without embedding rationale, functional usage, domain information or procedural information.
- ▶ Avoid circular reasoning.
- ▶ Use consistent terminology and structure for related definitions.

# Example Definition Syntax

27

- ▶ Use a phrase, not a sentence.

The name of the country where mail is delivered.

- ▶ Begin the definition by stating the representation class, such as:

The *name* of....

The *code* that represents....

The *text* that describes....

The *measure* of the....

The *number* assigned by...to identify....

The sum, dimension, capacity (*quantity*) of....

# The UML Model: Enterprise Architect Attributes

28

**Customer Attributes**

General | Detail | Constraints | Tags

Name:

Type:   ☐ Derived ☐ Static

Scope: **Private**  ☐ Property ☐ Const

Stereotype:

Containment: Not Specified

Alias:

Initial:

Notes:

Attributes

Name	Type	Scope
Account	CustomerAccount	Private
Notes	string	Public
Address	string	Private
City	string	Public
Country	Country	Public
CustomerID	string	Public
FirstName	string	Public
LastLogin	string	Public

# UML Fields for Attribute Metadata

29

Control	Description
Name	Attribute name
Type	Data type of attribute - select from the drop down list
Build button	Opens the <i>Select Attribute Type</i> dialog
Scope	Public/Protected/Private/Package
Stereotype	Optional Stereotype of the attribute
Containment	Containment type(by reference/value)
Derived	Indicates attribute is a calculated value
Static	Attribute is a static member
Property	Select automatic property creation
Const	Attribute is a constant
Alias	An optional alias for the attribute
Initial	An optional initial value
Notes	Free text notes
Attribute List	List of defined attributes. Select an attribute to make it current
Up/Down buttons	Use to change the order of attributes in the list
New	Create new attribute
Save	Save new attribute, or save modified details for existing attribute
Delete	Delete currently selected attribute

# UML Model Transformation

30

1. Create UML in Enterprise Architect and save it – be sure to fully populate with names and definitions for Classes and Attributes, Datatypes, and identifiers for related EVS terms.
2. Access class diagram and elements at UML Loader run-time.
3. Transform UML to caDSR Metadata using the automated tools.
4. Review the registration and make manually changes as needed to fully represent the model.

# Future Class and Class Attribute to OC, Property and DEC Mapping

31

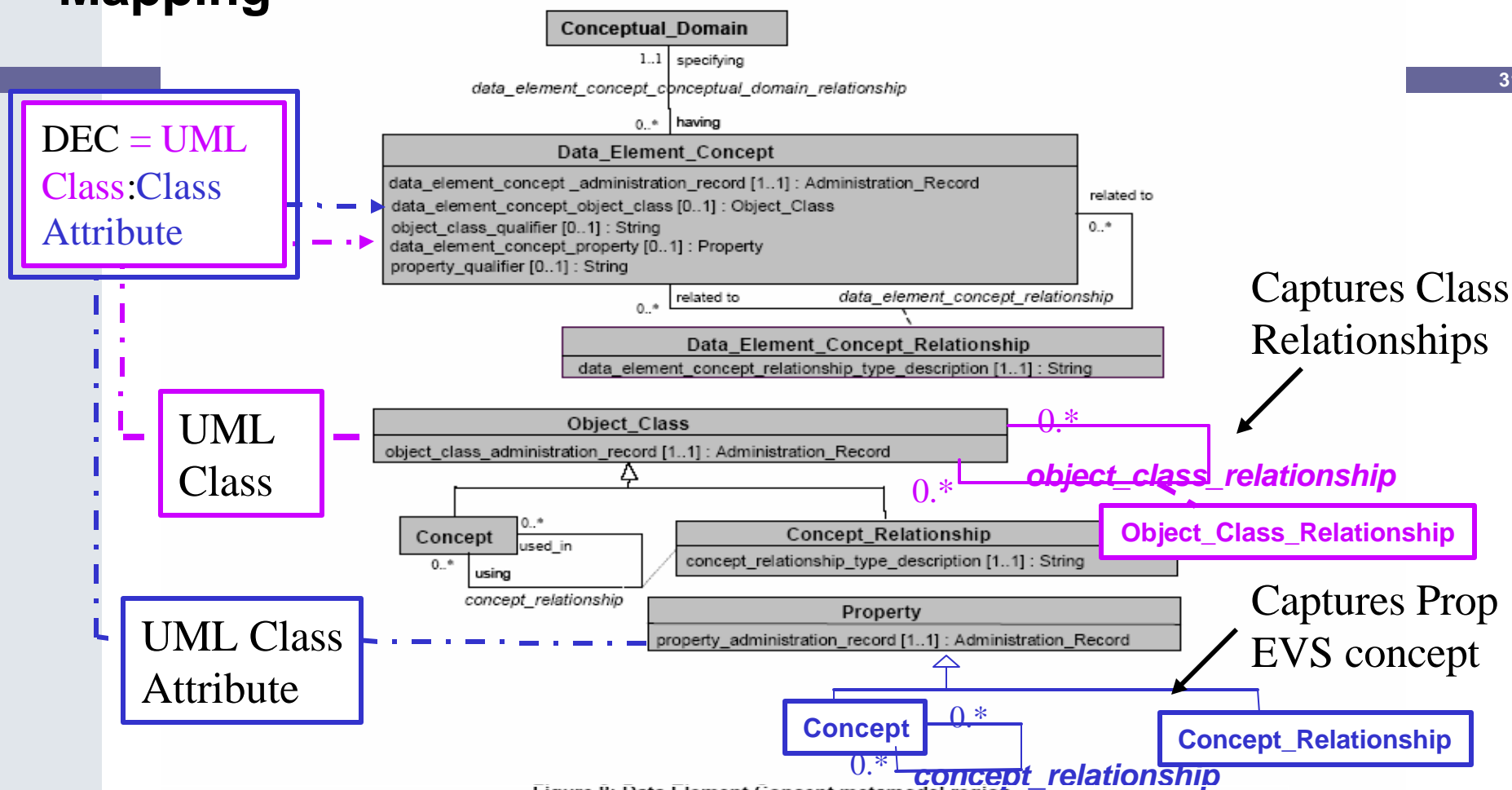


Figure 8: Data Element Concept metamodel region

e.g.

UML Class = Taxon; Class Attribute = commonName

caDSR OC = Taxon; Property = Common Name; DEC = Taxon Common Name

# Future Class Attribute = DE Mapping DEC now represents the Combination of the Class and Class Attribute see Taxon example below

ISO/IEC FDIS 11179-3:2002(E)

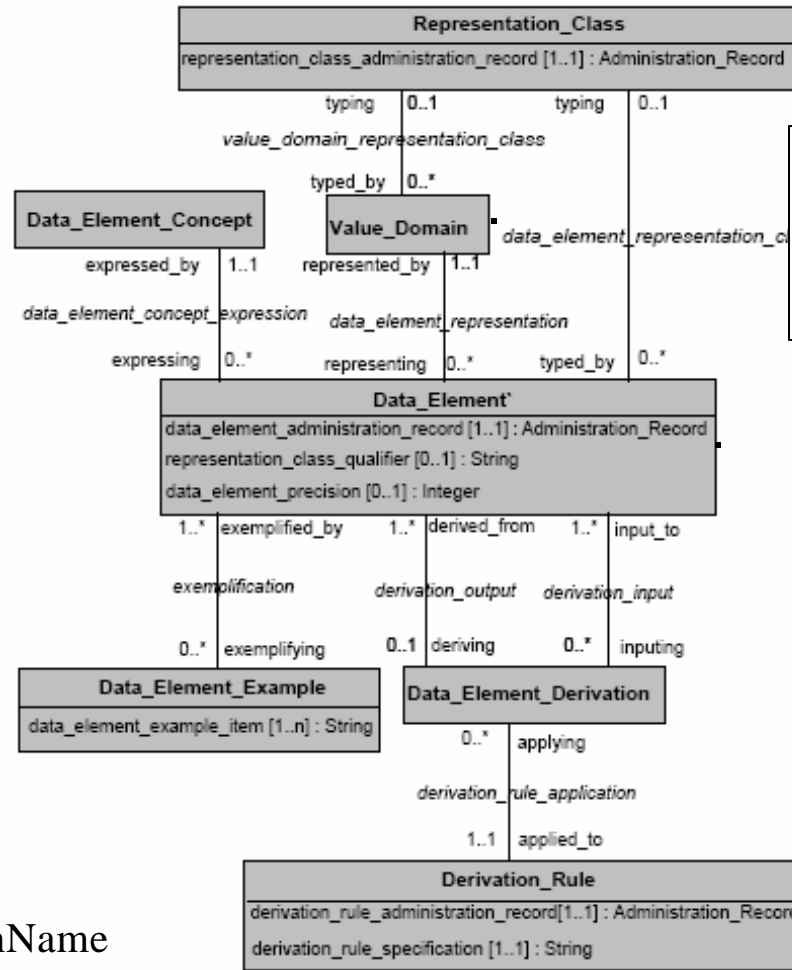
32

UML Class +  
Attribute = DEC  
Preferred Name

UML Class  
Attribute = DE  
Long Name

UML Class +  
Attribute =  
VD Datatype

UML Class  
Attribute Notes =  
DE Definition



e.g.

Class = Taxon

Attribute = commonName

DEC = Taxon Common Name

VD = Taxon Common Name

DE = commonName (caDSR naming convention not applied)



# Set UML Loader Defaults in Admin Tool

33

http://cadsr-prod.nci.nih.gov/pls/cadsr/meta\_ui.cmraccess - Microsoft Internet Explorer

**NATIONAL CANCER INSTITUTE** **caDSR** **caDSR Administration Tool** Release 2.1 [Logout](#) D:\WARZEL

**Metadata Browsing and Maintenance**

- [Data Element](#)
- [Data Element Concept](#)
- [Object Class](#)
- [Property](#)
- [Value Domain](#)
- [Representation](#)
- [Conceptual Domain](#)
- [Classification Scheme](#)

**Submissions/Registrations**

- [Stewards](#)
- [Submissions](#)
- [Registrations](#)

**System Administration**

- [Lookup Maintenance](#)
- [User Accounts](#)
- [User Groups](#)
- [Historical Data](#)
- [Security Maintenance](#)
- [Security Framework](#)

**Compliance Review Process**

- [Select CRF for Review](#)
- [Delete CRF](#)

**Protocol/Form/Template Browsing and Maintenance**

- [Protocol](#)
- [Form/Template](#)

**Other links**

- [Set UML Loader Defaults](#)
- [CDE Curation Tool](#)
- [CDE Browser](#)
- [CCRR](#)

Start | Inb... | caB... | ftp:... | NCI... | caB... | CD... | RE:... | can... | caB... | UM... | htt... | Internet | 10:30 PM

# Set UML Loader Defaults

34

cancer Data Standards Repository - Microsoft Internet Explorer

**NATIONAL CANCER INSTITUTE** **caDSR** **caDSR Administration Tool** Release 2.1 [Home](#) [Logout](#) [DWARZEN](#)

### Load Defaults

**Batch Id:**

**Context Name:**  [List](#)

**Context Version:**

**Version:**

**Workflow Status:**  [List](#)

**CD Preferred Name:**  [List](#) [New](#)

**CD Version:**

**CD Context Name:**

**CD Context Version:**

**VD Preferred Name:**  [List](#) [New](#)

**VD Version:**

**VD Context Name:**

**VD Context Version:**

**Reference Document Name:**

**Reference Document Type:**  [List](#)

**Reference Document Text:**

**Reference Document Name URL:**

**Classification Scheme:**  [List](#)

**Class scheme Item:**

Done Internet

Start In... C... Ft... N... C... C... R... C... C... U... ca... U... D... 10:36 PM

[Data Element](#)[Data Element Concept](#)[Permissible Values](#)[Classifications](#)[Usage](#)[Data Element Derivation](#)

## Selected Data Element

<b>Public ID:</b>	2178609
<b>Preferred Name:</b>	ProteinGeneInfold
<b>Long Name:</b>	geneInfold
<b>Document Text:</b>	
<b>Definition:</b>	The gene information identification for the protein object.
<b>Workflow Status:</b>	DRAFT MOD
<b>Version:</b>	2.1

## Data Element Concept Details

<b>Public ID:</b>	2178563
<b>Preferred Name:</b>	Protein
<b>Long Name:</b>	gov.nih.nci.caBIO.bean.Protein
<b>Definition:</b>	An object representation of a protein; provides access to the encoding gene via its GenBank ID, the taxon in which this instance of the protein occurs, and references to homologous proteins in other species.
<b>Context:</b>	caCORE
<b>Workflow Status:</b>	DRAFT MOD
<b>Version:</b>	2.1
<b>Conceptual Domain Preferred Name:</b>	BIOINFORMATICS
<b>Conceptual Domain Context Name:</b>	caCORE
<b>Conceptual Domain Version:</b>	1.0
<b>Object Class Preferred Name:</b>	
<b>Object Class Context:</b>	
<b>Object Class Version:</b>	
<b>Object Class Qualifier:</b>	
<b>Property Preferred Name:</b>	
<b>Property Context:</b>	
<b>Property Version:</b>	
<b>Property Qualifier:</b>	
<b>Origin:</b>	

# EVS Linkages to the ISO 11179 model

36

*EVS* is a terminology server that provides services for synonymy, mapping between vocabularies, hierarchical structures, Subconcepts, Superconcepts, Roles, Semantic type, etc.

- ▶ *EVS* is the source of metadata in the caDSR, including:

Data Element Concept Components:

- Object Class.
- Property.

Value Domain Representation Terms.

# Linking to EVS

37

- ▶ Objects and Property links will be made during curation or by during model development.
- ▶ Each term in EVS is assigned an identifying number. Different numbers are assigned in each EVS list; both the number and the number source must be kept.
- ▶ During automated model registration, terms can be added based on the name of a Class or Attribute. Appropriate EVS identifiers will be recorded as metadata in the caDSR.

# EVS Metathesaurus Example

38

C0033684: Protein



Amino Acid, Peptide, or Protein

Biologically Active Substance

## Protein Definitions



Source	Definition
<b>CSP2003</b>	linear polymers of alpha-L-aminoacids ranging in size from a few thousand to over 1 million daltons, capable of oligomerization, with specific functions dictated by aminoacid sequence and encoded genetically.
<b>MSH2004_2003_12_12</b>	Polymers of amino acids linked by peptide bonds. The specific sequence of amino acids determines the shape and function of the protein.
<b>NCI</b>	Any of a group of complex organic macromolecules that contain carbon, hydrogen, oxygen, nitrogen, and usually sulfur and are composed of one or more chains of amino acids. Proteins are fundamental components of all living cells and include many substances, such as enzymes, hormones, and antibodies, that are necessary for the proper functioning of an organism.
<b>NCI-GLOSS</b>	(PRO-teen) A molecule made up of amino acids that are needed for the body to function properly. Proteins are the basis of body structures such as skin and hair and of substances such as enzymes, cytokines, and antibodies.

# EVS Thesaurus Example

39

## Identifiers:

name	Protein
code	C17021

## Information about this concept:

Preferred_Name	Protein
Semantic_Type	Amino Acid, Peptide, or Protein
DEFINITION	CSP2000 linear polymers of alpha-L-aminoacids ranging in size from a few thousand to over 1 million daltons, capable of oligomerization, with specific functions dictated by aminoacid sequence and encoded genetically.
DEFINITION	MSH2001 Polymers of amino acids linked by peptide bonds. The specific sequence of amino acids determines the shape and function of the protein.
DEFINITION	NCI Any of a group of complex organic macromolecules that contain carbon, hydrogen, oxygen, nitrogen, and usually sulfur and are composed of one or more chains of amino acids. Proteins are fundamental components of all living cells and include many substances, such as enzymes, hormones, and antibodies, that are necessary for the proper functioning of an organism.

# Consuming CDEs – Download from Browser

40

Search Field(s): ALL  
Preferred Name  
Long Name  
Document Text

Alternate Name Type(s): ALL  
ABBREVIATION  
C3D Name

Permissible Value:

Value Domain:  [Clear](#)

Data Element Concept:  [Clear](#)

Version: ☒ Latest Version ☐ All Versions

Workflow Status: ALL  
APPRVD FOR TRIAL USE  
CMTE APPROVED  
CMTE SUBMTD

Public ID:

Classification:  [Clear](#)

Context Use: Owned By/Used By

Registration Status: ALL  
Application  
Candidate  
Qualified

[Search Data Elements](#) [Clear](#) [New Search](#)

[\[Download Data Elements to Excel\]](#) [\[Download Data Elements as XML\]](#)

[Add to CDE Cart](#) 1 - 2 of 2

<input type="checkbox"/>	Long Name	Document Text	Owned By	Used By Context	Registration Status	Workflow Status	Public ID	Version
<input type="checkbox"/>	<a href="#">proteinInfold</a>		caCORE			DRAFT MOD	2178636	2.1
<input type="checkbox"/>	<a href="#">proteinInfold</a>		caCORE			DRAFT MOD	2178610	2.1



# Consuming CDEs – caBio API

41



## caBIO Quick Start

caBIO provides its users with three different Application Programming Interfaces (APIs):

- [Java Interface](#) via the caBIO.jar file
- [Web Services Interface](#) via SOAP
- [HTTP/XML Interface](#) via a servlet

[http://ncicb.nci.nih.gov/core/caBIO/technical\\_resources/guides/quick\\_start](http://ncicb.nci.nih.gov/core/caBIO/technical_resources/guides/quick_start)

# Questions and Answers